Analiza i statystyka w nauczaniu maszynowym

Jacek Tabor

CZM, Będlewo 2014

Centrum Zastosowań Matematyki

Dane

Różne typy danych:

- dane „sztuczne”: generowane ze znanych rozkładów
- dane z repozytorium UCI
- dane „realne”: biologiczne, chemiczne, teksty, zdjęcia, etc

Zadania

Typowe zadania:

- klastrowanie: dzielimy na grupy (uczenie nienadzorowane)
- klasyfikacja: znamy przykładowe etykiety, chcemy przewidzieć (uczenie nadzorowane)

Klastrowanie: trudno ocenić jakość. Podstawowe metody: k-means, hierarchiczne, na bazie gęstości (EM), subspace clustering, etc.

Klasyfikacja: łatwo ocenić jakość. Podstawowe metody: SVM, Bayes, sieci neuronowe, regresja, ELM, drzewa decyzyjne i lasy losowe.

Cross-entropy clustering

Celem w tym referacie jest opowiedzenie o metodzie klastrowania opartej na entropii krzyżowej (cross-entropy clustering): CEC.

Połączenie metod stosowanych w metodzie k-means z podejściem EM (expectation maximization).

Literatura

Prace

TABOR, SPUREK:
Cross-entropy clustering, Pattern Recognition 2014

TABOR, MISZTAL:
Detection of elliptical shapes via cross-entropy clustering (IbPRIA 2013)

SPUREK, TABOR, ZAJĄC:
Detection of disk-like particles in electron microscopy images (CORES 2013)

ŚMIEJA, TABOR:
Image segmentation with use of cross-entropy clustering (CORES 2013)

Dostępne na mojej stronie `http://www.ii.uj.edu.pl/~tabor`.

Metoda k-means

Problem (optymalizacyjny)

Szukamy takiego rozbicia zbioru X na k podzbiorów $X_1, \ldots, X_k$, by minimalizować funkcję kosztu $\sum_{i=1}^{k} ss(X_i)$ gdzie ss (within cluster sum of squares)

$$ss(Y) := \sum_{y \in Y} \|y - m_Y\|^2,$$

a $m_Y = \frac{1}{|Y|} \sum_{y \in Y} y$ to średnia („środek ciężkości") zbioru Y.

Plusy: ciągle spotykana, bo szybka (umożliwia podejście Hartigana) i prosta w implementacji.

Wady: zależy od skalowania współrzędnych, nie znajduje ilości klastrów, klastry mniej więcej podobnej wielkości.

Podejście Lloyda do szukania optymalnego podziału

1. Wybieramy ze zbioru X w sposób losowy k punktów $m_1, \ldots, m_k$ (początkowe środki klastrów)
2. Kolejnym punktom z X przyporządkowujemy ten numer (indeks) środka dla którego $\|x - m_i\|^2$ jest najmniejsze
3. wyliczamy nowe środki, i o ile nastąpiła zmiana, wracamy do punktu drugiego

Metoda EM

Metoda EM (w najczęściej spotykanej postaci Gaussian Mixture Models). Stara się dopasować (biorąc pod uwagę funkcję największej wiarygodności) do danych X „mieszankę” postaci

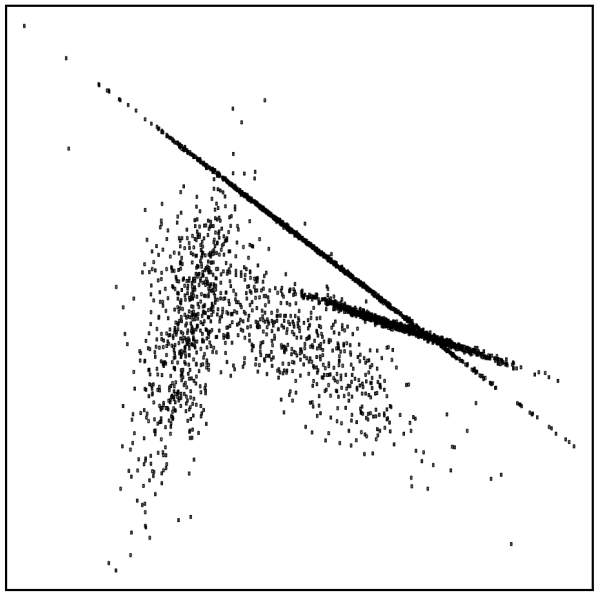
$$X \sim p_1 f_1 + \ldots + p_k f_k,$$

gdzie f_i należą do ustalonej wcześniej rodziny gęstości.

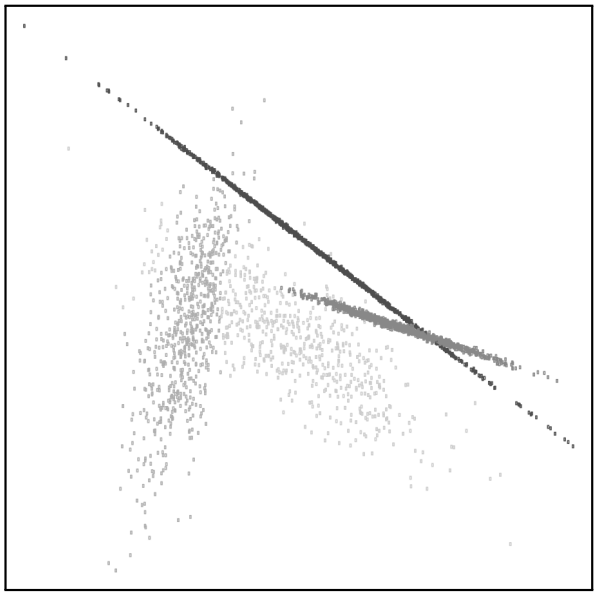
Plusy: nie zależy od skalowania, umożliwia dosyć dobre dopasowanie do danych, dokonuje estymacji gęstości, umożliwia użycie różnych typów gęstości (modeli Gaussowskich).

Wady: dosyć wolna (nie umożliwia podejścia Hartigana), nie znajduje automatycznie ilości klastrów, użycie nawet prostych modeli Gaussowskich wymaga skomplikowanej nieliniowej optymalizacji.

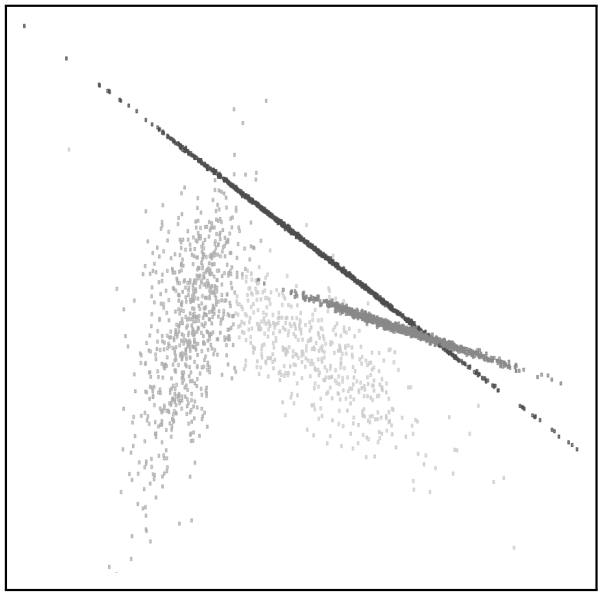
Przykład działania EM



(a) Dane pochodzące z mieszanki 4 gaussów.



(b) Klastrowanie EM z 4 gaussami.



(c) CEC startujący z początkowo 10 gaussami.

Rysunek: Porównanie klastrowania mieszanki 4 gaussów przy pomocy EM (z $k = 4$) z Gaussowskim CEC-em który zaczyna od $k = 10$ klastrów.

Co to CEC?

Dziedziczy najlepsze cechy k-means (prostota implementacji) z możliwościami EM (łatwość w użyciu różnych modeli Gaussowskich). Co więcej automatycznie redukuje niepotrzebne klastry.

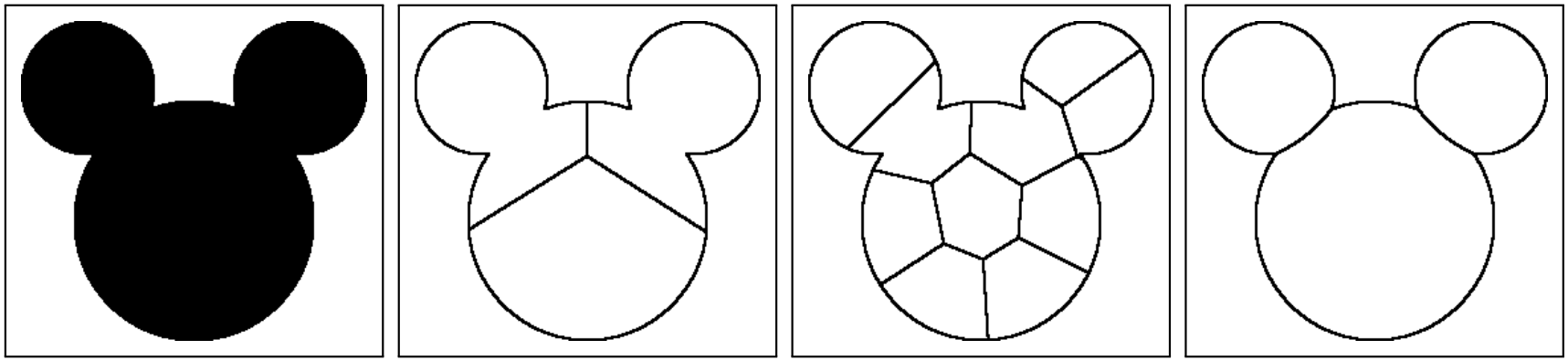
Działa podobnie jak EM, i stara się dopasować (biorąc pod uwagę funkcję największej wiarygodności) do danych X „mieszankę" postaci

$$X \sim \max(p_1 f_1, \ldots, p_k f_k),$$

gdzie f_i należą do ustalonej wcześniej rodziny gęstości.

Motywacja do powyższego wzoru pochodzi z teorii kodowania, entropii i paradygmatu MDLP (minimal description length principle).

Dlaczego CEC?

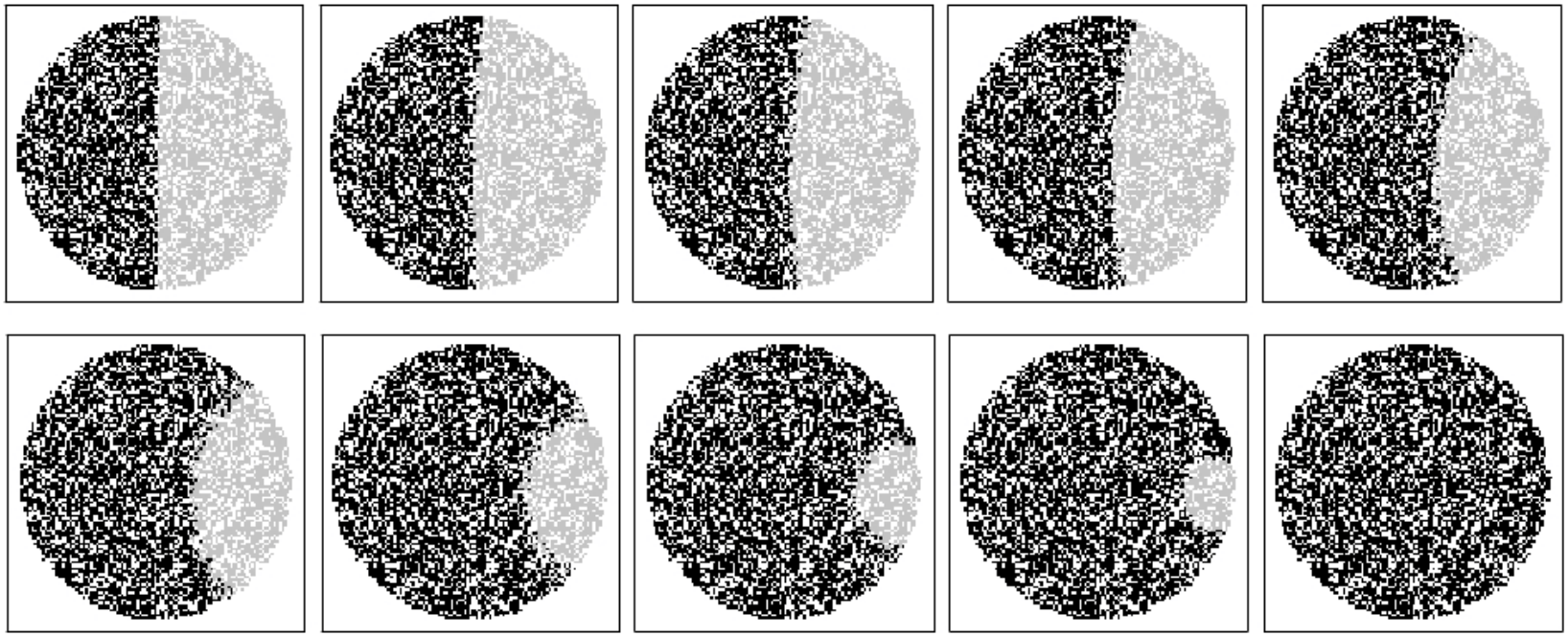


(a) zbiór typu Myszka Miki. (b) k-means z $k=3$. (c) k-means z $k=10$. (d) CEC start z 10-ma klastrami.

CEC:

- ładne klastry
- automatyczna redukcja ilości
- niezmiennicze ze względu na wybrane przekształcenia afiniczne
- struktura modułowa
- prostota zbliżona do k-means przy efektach EM

Redukcja klastrów



Rysunek: Redukcja klastrów przy sferyczym (radialnym) CEC-u.

W trakcie powstawania: Pakiet w R

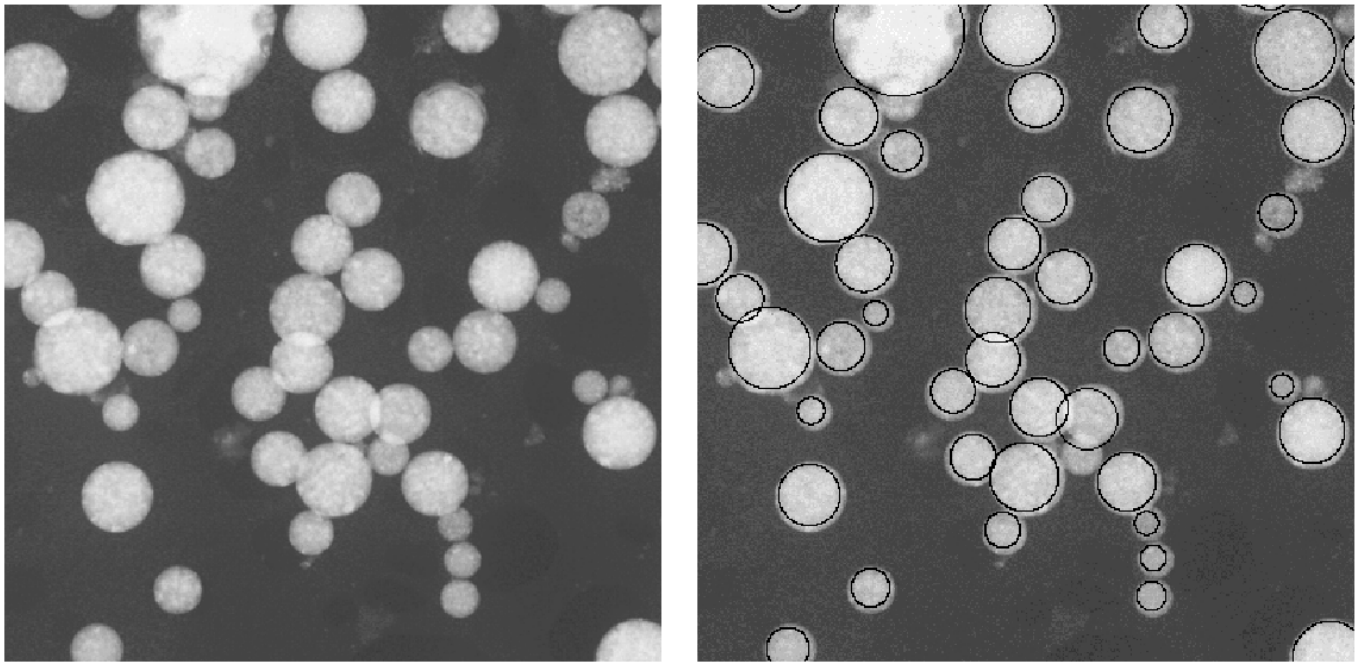
Dostępny już jest testowy pakiet do CEC. Pokażę implementację.

Dodatkowo, w naszej grupie roboczej GMUM (grupa metod uczenia maszynowego) `http://gmum.ii.uj.edu.pl` powstaje pakiet którego częścią bedzie CEC w R (planowane zakończenie: 2015).

Natomiast mając daną implementację k-means (zarówno metodą Lloyd'a lub Hartigana) można w ciągu dnia zaadaptować do CEC (najwięcej modyfikacji wymaga usuwanie „znikających" klastrów).

Wykrywanie kształtów kołowych/kulistych

Wyniki z pracy [Spurek,Tabor,Zając]:



(a) Zdjęcie.

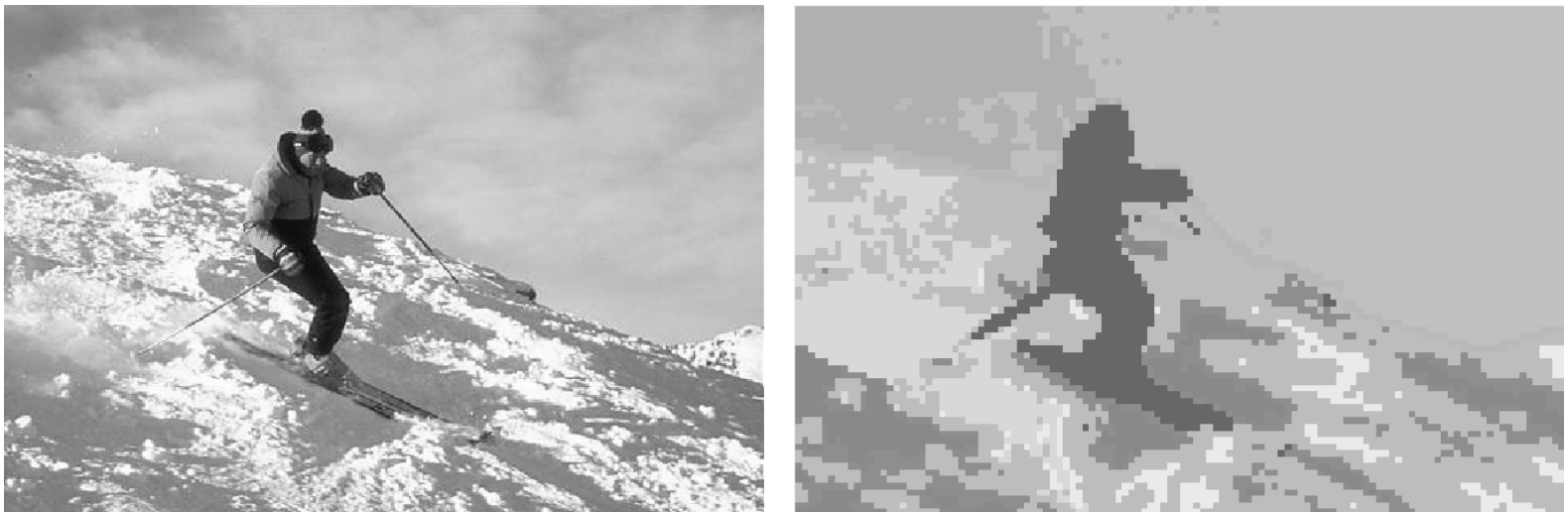
(b) Segmentacja uzyskana za pomocą sferycznego CEC.

Rysunek: Segmentacja sferyczna CEC cząstek nanopalladu.

Trudno byłoby wykonać przy pomocy EM, gdyż EM nie wykrywa ilości grup.

Segmentacja obrazów

Wyniki z pracy [Śmieja, Tabor]:



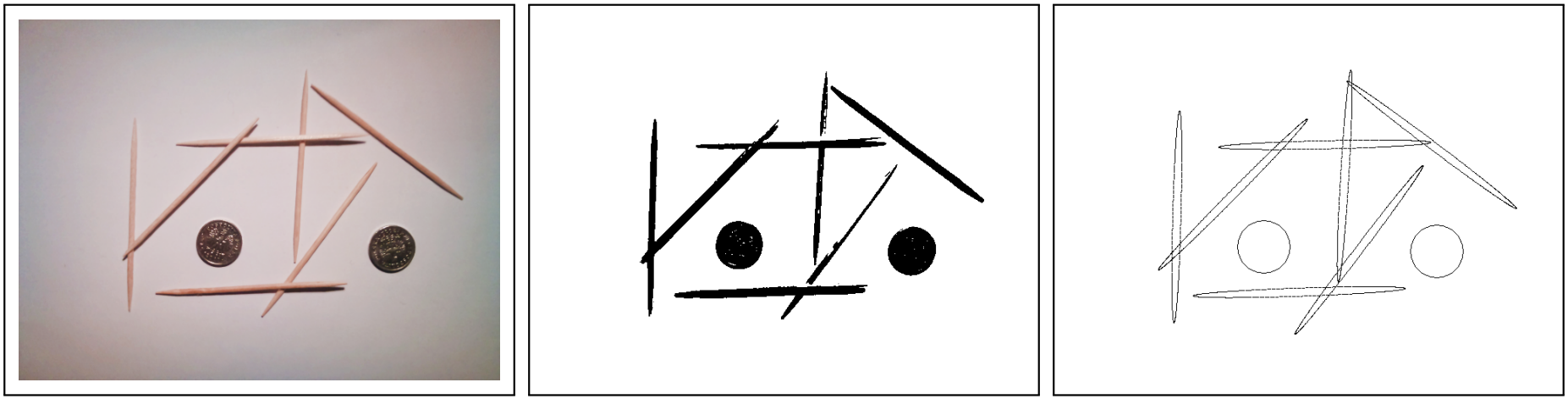
(a) Zdjęcie. (b) Segmentacja za pomocą CEC.

Rysunek: Segmentacja obrazów bazowana na CEC.

Zasadniczo niezmiennicze ze względu na afiniczne transformacje obrazu (w przeciwieństwie choćby do metody k-means).

Rozpoznawanie różnych typów kształtów eliptycznych

Wyniki z pracy [Tabor, Misztal]:



(a) wykałaczki i monety. (b) binaryzacja. (c) znalezione obiekty.

Rysunek: Rozpoznawanie obiektów.

Możliwość łatwego wyszukiwania ustalonych typów obiektów eliptycznych. Aby opisać jak to można robić, chciałbym przejść troszkę do teorii.

Rozkład normalny wielowymiarowy

CEC bazuje na rozkładach normalnych.

Wzór na $N(m, \Sigma)$ – wielowymiarowy rozkład normalny o wartości średniej m i kowariancji Σ w $\mathbb{R}^d$:

$$N(m, \Sigma) : x \to \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \exp(-\frac{1}{2}\|x - m\|_{\Sigma}^2),$$

gdzie $\|x - m\|_{\Sigma}^2$ to norma Mahalanobisa:

$$\|v\|_{\Sigma}^2 = v^T \Sigma^{-1} v.$$

Metoda Największej Wiarygodności

Jeżeli mamy rozkład normalny $N(m, \Sigma)$ i zestaw punktów $x_1, \ldots, x_n \in \mathbb{R}^d$. Wtedy wartość dopasowania tego rozkładu do danych jest dany z MNW (zmieniam znak na przeciwny, więc im miej tym lepsze dopasowanie):

$$-\sum_{i=1}^{n} \log(N(m, \Sigma)(x_i)) = \frac{nd}{2} \log(2\pi) + \frac{n}{2} \det(\Sigma) + \frac{1}{2} \sum_{i=1}^{n} \|x_i - m\|_{\Sigma}^2.$$

Kodowanie

Dualne patrzenie na MNW – za pomocą teorii kodowania (entropia różniczkowa).

Mając dany rozkład normalny $N(m, \Sigma)$, „długość kodu” dla punktu x wynosi

$$-\log(N(m, \Sigma))(x).$$

Mamy dane k-metod kodowania identyfikowanych z rozkładami normalnymi $N(m_i, \Sigma_i)$. Dodatkowo potrzebne są identyfikatory używanego algorytmu – $p_i \in (0, 1)$ które sumują się do jedynki.

Wtedy koszt kodowania punktu x za pomocą i-tej metody jest równy

$$-\log p_i - \log(N(m_i, \Sigma_i))(x).$$

CEC-klastrowanie: algorytm w $\mathbb{R}^d$

Algorytm postępowania:

1. ustalamy początkowe rozkłady $N(m_i, \Sigma_i)$ i ich identyfikatory p_i (m_i - losowo wybrane punkty z danych, $\Sigma_i = I$, $p_i = 1/k$)
2. idziemy po wszystkich punktach, i wrzucamy punkt do tej grupy do której mu najbliżej (w sensie długości kodu, patrz poprzednia strona) – powstają grupy $X_1, \ldots, X_k$
3. z każdej grupy wyliczamy średnią m_i, kowariancję Σ_i, $p_i = |X_i|/|X|$
4. tworzymy nowe metody klastrowania $N(m_i, \Sigma_i)$, p_i, i wracamy do punktu 2

Modele Gaussowskie

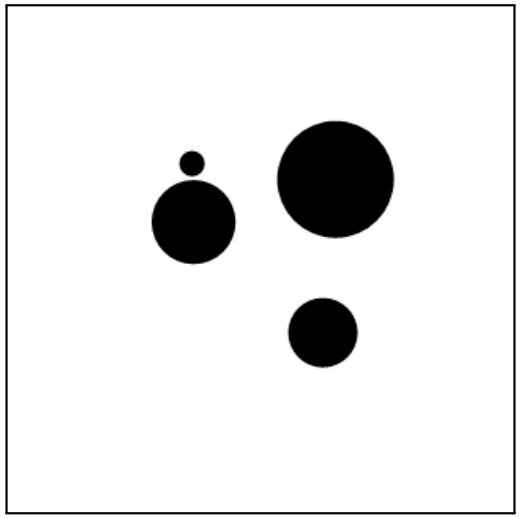
Biorąc odpowiednie podrodziny Gaussowskie możemy w różny sposób dopasowywać się do danych. Opisana wcześniej metoda znajduje najlepsze dopasowanie w klasie wszystkich rozkładów normalnych do zbioru X:

$$m = m_X,\ \Sigma = \Sigma_X.$$

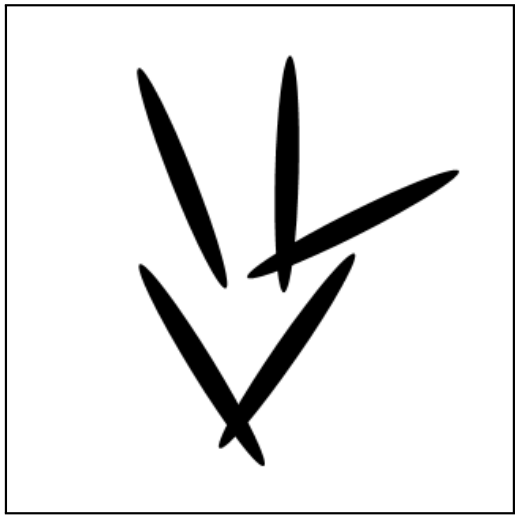
Czasami chcemy szukać najlepszego dopasowania w innych klasach *podrodzin Gaussowskich*. Najczęściej spotykaną jest rodzina rozkładów radialnych (sferycznych). Są to te rozkłady, które mają „symetrię radialną” (wartość gęstości zależy tylko od odległości od środka) – w konsekwencji oznacza to, że kowariancja musi być proporcjonalna do identyczności. Wtedy wzór na optymalne dopasowanie rozkładu radialnego do danych $X \subset \mathbb{R}^d$ to

$$m = m_X,\ \Sigma = \frac{\mathrm{tr}\Sigma_X}{d} I.$$

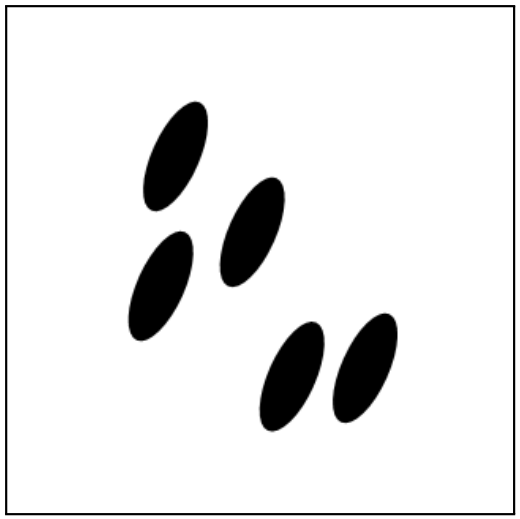
(a) Wszystkie gaussy (Cov dowolna).



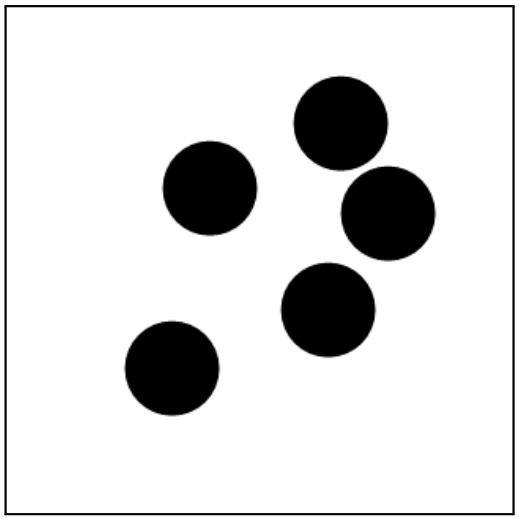
(b) Gaussy radialne (Cov proporcjonalna do I).



(c) Cov o ustalonych wartościach własnych.



(d) Zafiksowana kowariancja.



(e) Cov$= I$.